University of Zurich[UZH]

Zurich Open Repository and Archive
University of Zurich
University Library
Strickhofstrasse 39
CH-8057 Zurich
www.zora.uzh.ch

Year: 2021

On the reproducibility of in vivo temporal signal-to-noise ratio and its utility as a predictor of subject-level t-values in a functional magnetic resonance imaging study

Gobbi, Susanna ; Lee, Yoojin ; Homolya, István ; Tobler, Philippe N ; Hare, Todd A ; Nagy, Zoltan

Abstract: The aim of this study was to evaluate the reproducibility of voxel-wise temporal signal-to-noise ratio (tSNR) on repeated scans across runs, sessions, and days. A group of 21 participants was scanned 16 times (4 runs per session, 2 sessions per day, 2 separate days) in a functional magnetic resonance imaging (fMRI) study on a 3T Philips Achieva scanner. For each run, we calculated t-value and tSNR maps. To ascertain that the results were not specific to the scanner, one volunteer was scanned with four fMRI runs in a single session on the above 3T Philips scanner as well as a 3T Siemens Prisma scanner. The coefficient of variation of voxel-wise tSNR across the 16 repeats was up to 25%, while the range relative to the mean of all observations was up to 80%. The voxel-wise variability of tSNR on the two different scanners was similar, indicating a general issue. Despite its use in evaluating the quality of fMRI data, we found only a weak relationship between tSNR and t-values. There is very high variability in voxel-wise tSNR, which should be considered while planning future studies that aim to identify small and focal fMRI effects or the benefits of incremental improvement in methods.

DOI: https://doi.org/10.1002/ima.22617

Posted at the Zurich Open Repository and Archive, University of Zurich
ZORA URL: https://doi.org/10.5167/uzh-205637
Journal Article
Published Version

The following work is licensed under a Creative Commons: Attribution-NonCommercial-NoDerivatives 4.0 International (CC BY-NC-ND 4.0) License.

Originally published at:
Gobbi, Susanna; Lee, Yoojin; Homolya, István; Tobler, Philippe N; Hare, Todd A; Nagy, Zoltan (2021). On the reproducibility of in vivo temporal signal-to-noise ratio and its utility as a predictor of subject-level t-values in a functional magnetic resonance imaging study. International Journal of Imaging Systems and Technology, 31(4):1849-1860.
DOI: https://doi.org/10.1002/ima.22617

Check for updates

Received: 26 November 2020 | Revised: 26 April 2021 | Accepted: 27 May 2021

DOI: 10.1002/ima.22617

RESEARCH ARTICLE

WILEY

On the reproducibility of in vivo temporal signal-to-noise ratio and its utility as a predictor of subject-level t-values in a functional magnetic resonance imaging study

Susanna Gobbi[1] | **Yoojin Lee**[1] | **István Homolya**[2] | **Philippe N. Tobler**[1] | **Todd A. Hare**[1] | **Zoltan Nagy**[1]

[1]Laboratory for Social and Neural Systems Research, University of Zurich, Zurich, Switzerland

[2]Brain Imaging Centre, Research Centre for Natural Sciences, Budapest, Hungary

Correspondence
Susanna Gobbi, Department of Economics, Center for Neuroeconomics, University of Zurich, Blümlisalpstrasse 10, Room BLU-104, 8006 Zurich, Switzerland.
Email: susanna.gobbi@econ.uzh.ch

Funding information
Baugarten Stiftung; Foundation for Nutrition Research; Foundation for Scientific Research University of Zurich; Hungarian Brain Research Program (Nemzeti Agykutatási Program), Grant/Award Number: 2017-1.2.1-NKP-2017-0; Marlene-Porsche Foundation; Philhuman Foundation; Schweizerischer Nationalfonds zur Förderung der Wissenschaftlichen Forschung, Grant/Award Numbers: 100014 165884, 32003B_166566, PP00P1 150739; Seventh Framework Programme, Grant/Award Number: 607310; Zurich Center for Integrative Human Physiology

Abstract

The aim of this study was to evaluate the reproducibility of voxel-wise temporal signal-to-noise ratio (tSNR) on repeated scans across runs, sessions, and days. A group of 21 participants was scanned 16 times (4 runs per session, 2 sessions per day, 2 separate days) in a functional magnetic resonance imaging (fMRI) study on a 3T Philips Achieva scanner. For each run, we calculated t-value and tSNR maps. To ascertain that the results were not specific to the scanner, one volunteer was scanned with four fMRI runs in a single session on the above 3T Philips scanner as well as a 3T Siemens Prisma scanner. The coefficient of variation of voxel-wise tSNR across the 16 repeats was up to 25%, while the range relative to the mean of all observations was up to 80%. The voxel-wise variability of tSNR on the two different scanners was similar, indicating a general issue. Despite its use in evaluating the quality of fMRI data, we found only a weak relationship between tSNR and t-values. There is very high variability in voxel-wise tSNR, which should be considered while planning future studies that aim to identify small and focal fMRI effects or the benefits of incremental improvement in methods.

KEYWORDS

fMRI, quality assurance, reliability, reproducibility, temporal SNR

1 | INTRODUCTION

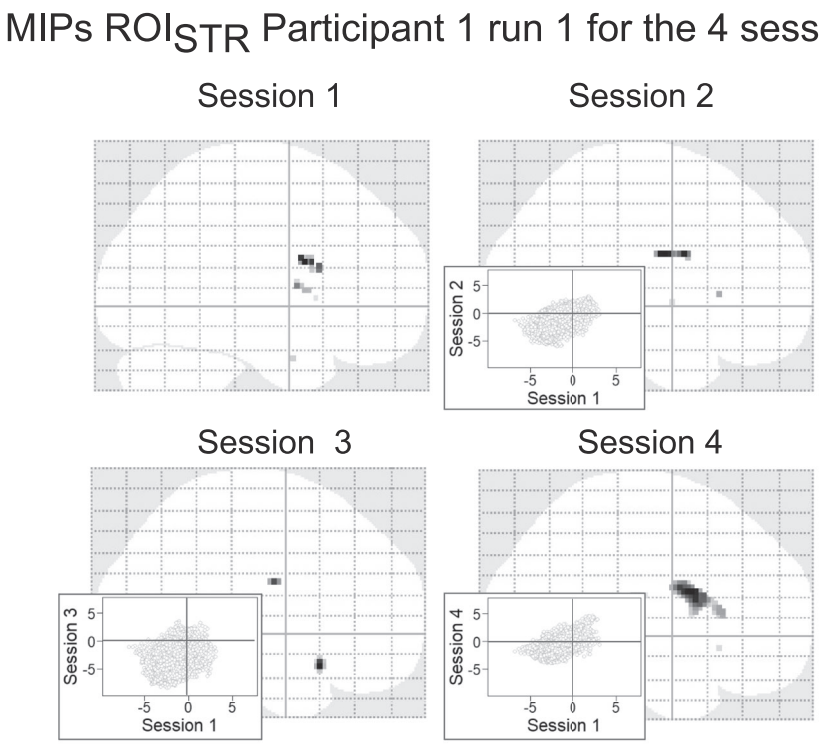
The reproducibility of scientific inquiry is becoming an increasingly hot topic.[1,2] A recent survey found that an alarming proportion (70% of 1576) of scientists have tried, but failed, to reproduce their own or others' results.[3] Reproducibility has been a focal discussion point also in (functional) magnetic resonance imaging (MRI)[4–9] (see also the 2019 special issue of Neuroimage on "Reproducibility in Neuroimaging").

The reliability of functional MRI (fMRI) has been repeatedly tested[7,10–16] and its best practice widely discussed.[8,17–21]

This is an open access article under the terms of the Creative Commons Attribution-NonCommercial-NoDerivs License, which permits use and distribution in any medium, provided the original work is properly cited, the use is non-commercial and no modifications or adaptations are made.
© 2021 The Authors. *International Journal of Imaging Systems and Technology* published by Wiley Periodicals LLC.

Some investigators have concluded that fMRI results were sufficiently reproducible,[10,11,15] while others have concluded that the state-of-the-art needed improvement.[7,13] However, to make effective improvements to scanner design, acquisition methods, or image processing, we need a reliable outcome measure of fMRI data quality for evaluating the potential benefits of the new methods.

One line of thinking advocates t-values as such a measure and previous efforts used t-values to investigate reproducibility of fMRI results. Still, t-value maps depend on the model choice at both single subject and group level and show poor test retest reliability[13,14] even for large-scale datasets like the Human Connectome Project.[7] Similar to those reports, we found large



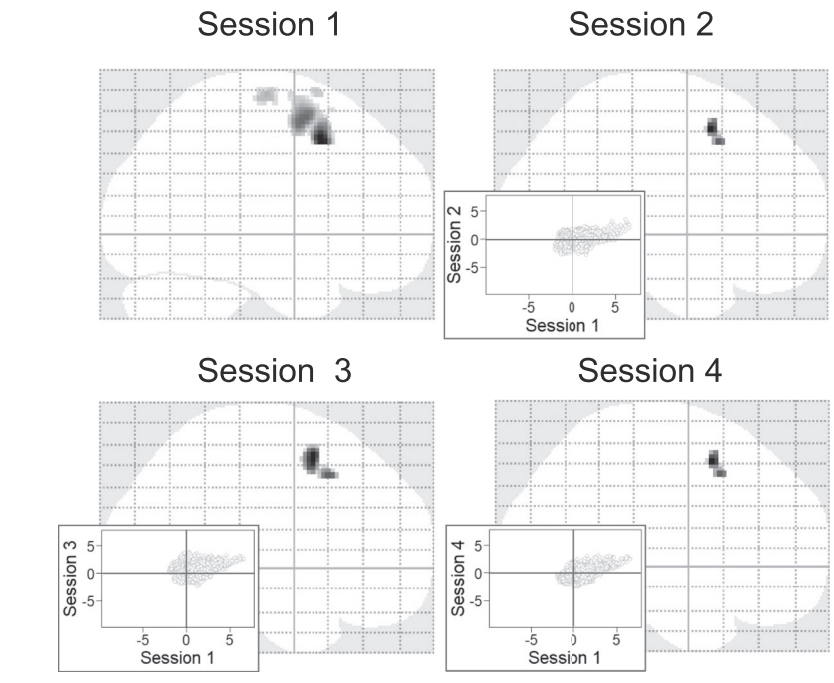


FIGURE 1 MIP examples from different participants and regions of interest (ROIs). The t-value maps are outputs of the first level analysis in SPM12. All four sessions are provided in both ROIs but focus on a different functional magnetic resonance imaging (fMRI) run (i.e., Run 1 or Run 2) for each of the two participants. The t-value threshold was 2.33 with $p < 0.01$ uncorrected. The scatter plots inset in the t-value maps for Sessions 2–4 show the correlations between voxel-wise t-values in that session and Session 1. The plots show the relationship between t-values in all voxels within the ROIs without applying any threshold for t-value magnitude

differences when comparing t-maps of fMRI runs from different sessions in our study (Figure 1).

Thus, we need a more reliable measure of data quality for fMRI studies. Here, we discuss and investigate the possibility of using the temporal signal-to-noise ratio (tSNR) to this aim. Often tSNR is measured from images of a water or gel phantom.[22,23] However, in vivo data are also necessary—for example, to investigate physiological noise,[24] compare acquisition parameter settings,[25] evaluate MRI sequence variants,[26] or optimize multiecho fMRI acquisitions.[27]

Let us consider a thought experiment in which the flip angle of an excitation pulse is varied for an fMRI study with TR = 0.5 s, which is common with multiband imaging.[28] Although a theoretically straightforward calculation provides the flip angle (i.e., the Ernst angle) for optimal SNR, the practical implementation is not trivial, because it requires a reliable estimate of T1 relaxation time. Apart from the challenges of obtaining reliable voxel-wise in vivo T1 maps, we must also consider that this measure is spatially variable across the cortex. At first, a practical approach seems reasonable. Assuming that T1 for gray matter (GM) is between 1.0 and 1.5 s, which would lead to Ernst angles between 42 and 53°, we may plan repeated experiments while incrementally changing the excitation flip angle. Then measuring the tSNR for the acquired data in the relevant area of the cortex should identify the optimal flip angle. This optimization process, however, is only reasonable if we know the voxel-wise reproducibility of tSNR so that an appropriate power calculation could be carried out prior to such an experiment.

Despite its wide use as a measure of quality, the voxel-wise reproducibility of tSNR has not been investigated thoroughly—especially for in vivo settings. However, without a reliable estimate of its reproducibility, it can be error-prone to utilize tSNR for identifying an improvement in methods or data quality. Furthermore, underestimating the variability of tSNR (i.e., overestimating statistical power) in a study that uses a small sample size (as is common in neuroimaging studies) leads to unreliable statistical inference.[29]

Therefore, the main aims of this study were to assess the voxel-wise reproducibility of in vivo tSNR and the relationship between tSNR and subject-level t-values in an fMRI study.

2 | METHODS

2.1 | Data acquisition

2.1.1 | Participants

Two different data sets were acquired with ethics approval of the relevant governing bodies in Switzerland and Hungary. Each participant in both data sets signed a written informed consent form. DataSet_1_Group included 21 female participants with a mean age of 25 years, scanned in four different sessions: on two different days during two cycle phases (preovulatory/postovulatory) and two different times of each day with different satiety states (fasted/fed). DataSet_2_Compare involved one adult male participant (44 years), scanned in a single session on two different scanners.

2.1.2 | Imaging protocol

DataSet_1_Group was collected on a 3T Philips Achieva MRI scanner (Philips Healthcare, Best, The Netherlands) with the vendor's eight-channel (SENSE) head coil.

Each scanning session contained four fMRI runs of 5.5 minutes each. The T2*-weighted echo-planar images were acquired with two, slightly different repetition times (TR) of 2370 or 2381 ms (because of a scanner software upgrade), echo time (TE) of 30 ms, flip angle of 90°, 40 oblique slices in ascending order with 0.5 mm gap, angulation of −20° along the LR axis, 3 mm isotropic voxel size and field of view (FOV) of [240 mm (AP), 139.50 mm (FH), 240 mm (RL)]. Each session also included a dual gradient-echo B_0 field map with TR/TE1/TE2 of 474/4.3/7.4 ms, flip angle of 44°, a total of 42 oblique slices, 3 mm isotropic voxels and FOV of [240 mm (AP), 156.75 mm (FH), 240 mm (RL)]. One of the four sessions also included a 3D T1-weighted anatomical sequence (3D FFE T1) with 1 mm^3 voxels.

DataSet_2_Compare included two separate sessions with four fMRI runs each but acquired on two different scanners. One session used the same Philips scanner and imaging protocol as that for DataSet_1_Group, the other session used an analogous imaging protocol on a 3T Siemens Prisma scanner with the vendor's 32ch receive-only head coil (Siemens Healthcare, Erlangen, Germany). The participant was awake with eyes closed during the measurements.

Importantly, all the Philips data were exported with an inverse scaling to ensure that the voxel signal intensities of any run/session/day would be comparable.[30]

2.1.3 | Taste task

During the acquisition of DataSet_1_Group, participants identified different liquids or rated either their pleasantness or intensity (for details, see Reference 31). The taste stimuli consisted of chocolate and strawberry milkshakes with different caloric content (low/high) and a neutral liquid (artificial saliva). Each fMRI run comprised

30 trials (drops), 10 trials per question type (pleasantness, intensity, and identity).

In each trial, participants first saw a drop cue and after 0.5 s received one drop of milkshake (2 s). After a blank screen (0.5 s), a question was displayed for 3.5 s. Finally, participants were instructed to swallow (1.5 s). Intertrial intervals lasted 3 s on average. For the present study, we focus on liquid delivery because it constitutes the main event of our taste task.

2.2 | Data analysis

All analyses were performed in MATLAB (MathWorks, Natick, MA) with the SPM12 software package[32] as well as custom-made MATLAB scripts. Three different analysis pipelines were implemented to establish the voxel-wise reproducibility of tSNR maps in the participant's native space as well as in standard MNI space and to investigate the predictive value of tSNR toward the subject-level statistical results of the fMRI study. Both datasets and all three processing pipelines were resampled using the third degree B-spline interpolation in SPM.

2.2.1 | Preprocessing

Analysis_1_tSNR_Nat: Describes the calculation of tSNR maps without general linear model fitting in the native space of individual participant brains

For DataSet_1_Group, the four runs of each session were realigned to the first volume of the first run, corrected for B_o field distortions, high-pass filtered with a cut-off frequency of 1/120 Hz and finally aligned to the participant's anatomical image via the mean of the realigned and filtered time series. These steps were performed using SPM12. For each of the 16 runs per participant, a voxel-wise tSNR map was calculated by dividing the temporal mean (tMean) by the temporal standard deviation (tSTD) within each voxel. As measures of reproducibility, we calculated the following voxel-wise maps for the tSNR, tMean, and tSTD maps:

a. Coefficient of variation (CoV) across the 16 runs
b. Range over mean (RoM) across the

- sixteen repeats (RoMAll)
- four repeats of each session (RoMRep)
- first run of each session on Day 1 (RoMSes)
- first run of the first session of the 2 days (RoMDay)
- first run of each of the four sessions (RoM1st)

The RoM maps were calculated voxel-wise as

$$\frac{\max\lfloor\text{of tSNR observations}\rfloor - \min\lfloor\text{of tSNR observations}\rfloor}{\text{mean}\lfloor\text{of tSNR observations}\rfloor}$$

because some of the comparisons were made over as few as two repeats where extracting variance for CoV calculations would be unreliable.

As part of the normalization process, the T1-weighted anatomical image was also segmented into separate tissue types[33] and the GM segment was thresholded at 0.8. The GM mask was used for masking the results for display in Figures 2 and 3. In Figure 2(B) where the group-level average results will be given, the GM mask was created by thresholding the MNI tissue probability map for GM at 0.5. The thresholds were chosen to provide GM masks that provide comparable visualization of the individual and group-averaged results.

DataSet_2_Compare, which consists of a single session of four fMRI repeats from two different scanners, was processed identically but, only the RoMRep (i.e., the variability across different fMRI runs within a single session) was calculated for both of the scanners.

Analysis_2_tSNR_MNI: Describes the calculation of tSNR maps after the data were warped to MNI space without general linear model fitting

DataSet_1_Group was treated as in Analysis_1 except that tSNR, tMean, and tSTD images were warped to the standard MNI space via the anatomical image of the participant before calculating the reliability measures for tSNR, tMean and tSTD. Warping to a standard space is typical for fMRI studies. Therefore, we ran Analysis_2_tSNR_MNI to determine what effects the warping procedures might have on the reproducibility of tSNR and to test its relationship to subject-level t-values (see next paragraph).

Analysis_3_fMRI_MNI: Describes the calculation of first-level t-value maps in MNI space

Each of the 16 realigned, B_o field corrected and high-pass filtered time series for each volunteer in DataSet_1_Group was smoothed using a 6-mm FWHM isotropic Gaussian kernel. Then, we corrected for physiological noise collected at the scanner via RETROICOR using Fourier expansions of different orders for the estimated phases of cardiac pulsation (third order), respiration (fourth order), and cardiorespiratory interactions (first order). We used the MATLAB PhysIO Toolbox[34] to create the corresponding confound regressors for the subject-level fMRI analysis (details in next paragraph and Reference 31). One of the four runs was excluded for three participants because of multiple instances of volume-to-volume displacement greater than 2 mm.

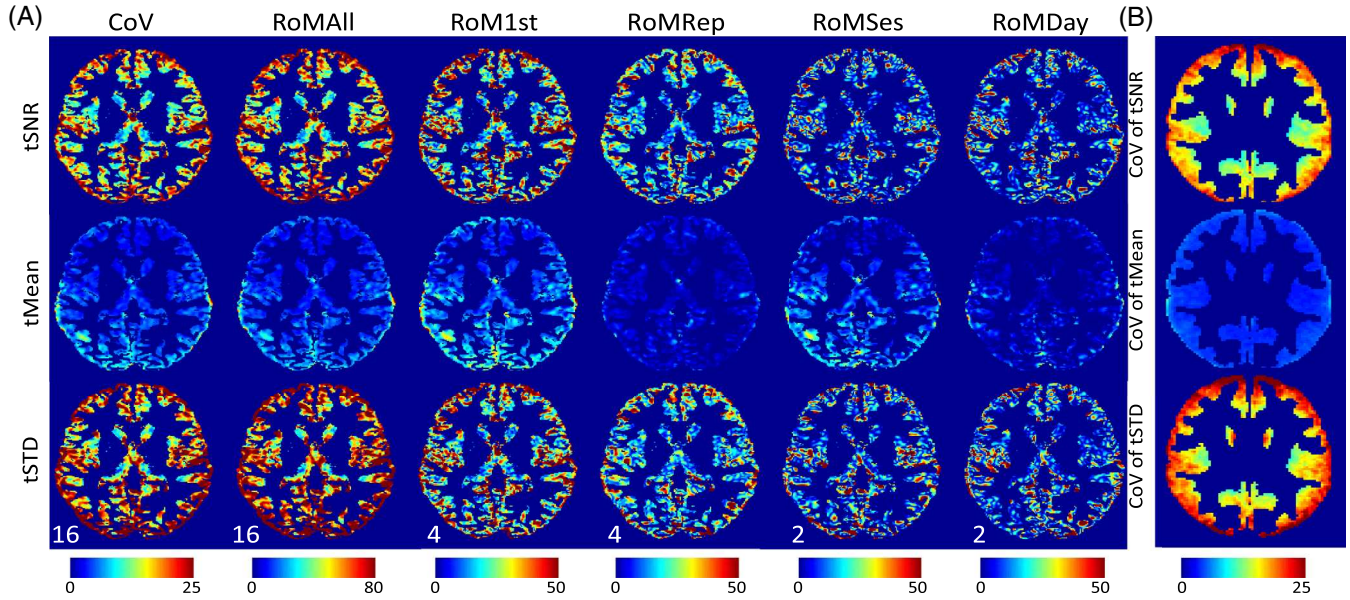


FIGURE 2 Variability measures for a single participant (native space) and the group-level average of the coefficient of variation (CoV) (MNI space). (A) Reproducibility measures of temporal signal-to-noise ratio (tSNR; top row), temporal mean (tMean; middle row) and temporal standard deviation (tSTD; bottom row) for a representative participant (#6). The tMean is a remarkably stable measure across repetitions, sessions and even days. Therefore, variability in tSNR (i.e., tMean/tSTD) is clearly a result of variability in tSTD. Note that the different columns have unique color scales and the CoV and range over mean (RoM) values are calculated across a different number of observations (indicated by a white number in the bottom left corner of each column) as appropriate in each case. (B) CoV for tSNR, tMean, and tSTD after normalization to MNI space and averaging across the 21 participants confirms the observation from native space for a single individual in Part (A) [Color figure can be viewed at wileyonlinelibrary.com]

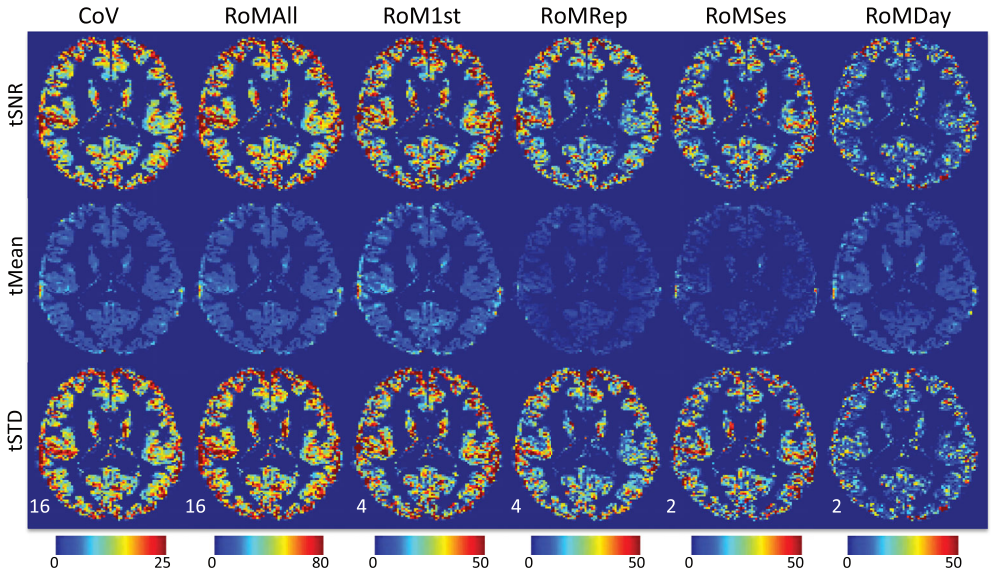


FIGURE 3 Variability measures of single participant (MNI space). Reproducibility measures of temporal signal-to-noise ratio (tSNR; top row), temporal mean (tMean; middle row), and temporal standard deviation (tSTD; bottom row) for a representative participant (#1). The findings in MNI space match those in native space. Accordingly, tMean is a much more stable measure and variability in tSNR (i.e., tMean/tSTD) is mainly a result of variability in tSTD. Note that the different columns have unique color scales and the range over mean (RoM) values are calculated across a different number of observations (indicated by a white number in the bottom left corner of each column) as appropriate in each case [Color figure can be viewed at wileyonlinelibrary.com]

Namely, the participants showed head motion higher than 2 mm with both drifts and spikes for more than 25% of the volumes in the run. The possible relation between tSNR and subject level t-values was investigated in relevant regions of interest (ROIs) (described below).

2.2.2 | Individual-level BOLD effect

Utilizing general linear models (GLMs) in SPM12, a design matrix with four runs was created with the BOLD time-series data in each voxel as the dependent variable. To focus on the brain responses at the time stimuli (visual images + drops of milkshake) were presented, we modeled the milkshake delivery in each trial as boxcar functions with duration 3.5 s and included 24 regressors to control for physiological noise. We convoluted the task onset regressors with the canonical hemodynamic response function before entering them into the GLM.

2.2.3 | Regions of interest

The relationship between tSNR and t-values was examined in two anatomical ROIs: the striatum (striatum ROI) and the supplementary motor area (SMA ROI) derived from the AAL3 atlas in SPM.[35] We selected the regions that consistently activated when people see pictures of food or taste food.[36,37] Moreover, we ran a group-level analysis in these ROIs to confirm that there was statistically significant activation in most of the voxels ($p < 0.05$ voxel-wise FWE-corrected).

Then, from each ROI, the mean tSNR and the mean t-value were extracted and evaluated using Pearson correlations within each participant or mixed-effects linear regressions across all participants and runs. Separate correlations were made across (a) the four runs in each session and (b) all of the 16 runs from each participant. The mixed-effects regressions were computed using the lme4 package in R.[38] These regressions sought to explain variability in t-values across individuals and runs as a function of a constant intercept and the tSNR in the corresponding run. The regressions included participant-specific intercepts and slopes for tSNR. In order to test the added value of including tSNR in the regression models, we compared the absolute value of the errors between model-predicted and observed t-values from a model including tSNR as a regressor and a model that included only an intercept term that captured the mean t-value. The absolute error values from each model were used to construct the histograms in Figure 5.

2.2.4 | Additional outcome measures

Additional analyses checked for confounds of demographic data such as age and body mass index (BMI). We correlated this information for each participant with the tSNR signal extracted from the two ROIs, averaging across runs and sessions. Possible correlations between the time laps between Day_1 and Day_2 and the difference in the standard deviation of the mean tSNR calculated in each day were also tested. Moreover, we investigated the variability in the mean tSNR across runs and sessions to look for potential noise given by time-of-day or subject-related physiological effects.

3 | RESULTS

3.1 | tSNR reproducibility in native space

Analysis_1_tSNR_Nat on DataSet_1_Group examines variability in the participants' native space (Figure 2(A)). Both the CoV and RoM measures of tSNR show a large amount of variability. Notably, the RoM across the 16 repeats (RoMAll) indicates that repeated scanning of the same individual using an identical sequence can produce tSNR values that are up to 80% different relative to

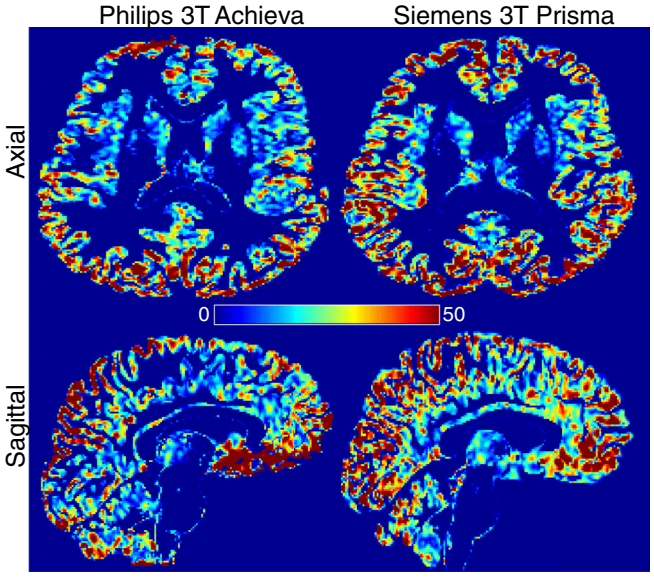


FIGURE 4 Data from two scanners confirm that temporal signal-to-noise ratio (tSNR) variability is not scanner specific. The RoMRep measure for the single participant in DataSet_2_Compare and from a single session on two separate scanner models results in similar levels of variability in voxel-wise tSNR. Both axial (top) and sagittal (bottom) orientations are shown [Color figure can be viewed at wileyonlinelibrary.com]

the mean of the observations. Therefore, using tSNR as an outcome measure to evaluate the benefits of new hardware or pulse sequence variants would be reasonable only for very large effect sizes and/or very large samples.

The bottom two rows of Figure 2(A) display the same measures of variability for the component parts of tSNR, and reveal that the surprisingly large variability in tSNR is mainly due to a fluctuation in tSTD from run to run, while tMean of the 16 runs remains relatively stable.

3.2 | tSNR reproducibility in MNI space

Calculating the CoV and RoM measures for tSNR, tMean, and tSTD after warping the images to MNI space (i.e., Analysis_2_tSNR_MNI) produced similar results both on an individual level (Figure 3) and when averaged across the entire DataSet_1_Group (Figure 2(B)).

3.3 | Scanner comparison

Running the Analysis_1_tSNR_Nat pipeline on DataSet_2_Compare confirms that the variability in DataSet_1_Group is not a specific problem of the scanner type or this particular scanner installation (Figure 4). The RoMRep values were computed across the four repeated runs of the same fMRI sequence and show similar variability for both 3T scanners, and can be up to 50%.

3.4 | tSNR vs t-value in ROIs

We did not observe a consistent relationship between mean tSNR and mean t-value in two relevant ROIs where significant activation was seen at the group-level during the milkshake taste task (Figure 5). Within neither session (first row) nor participant (second row) was there a consistent positive correlation that would indicate that

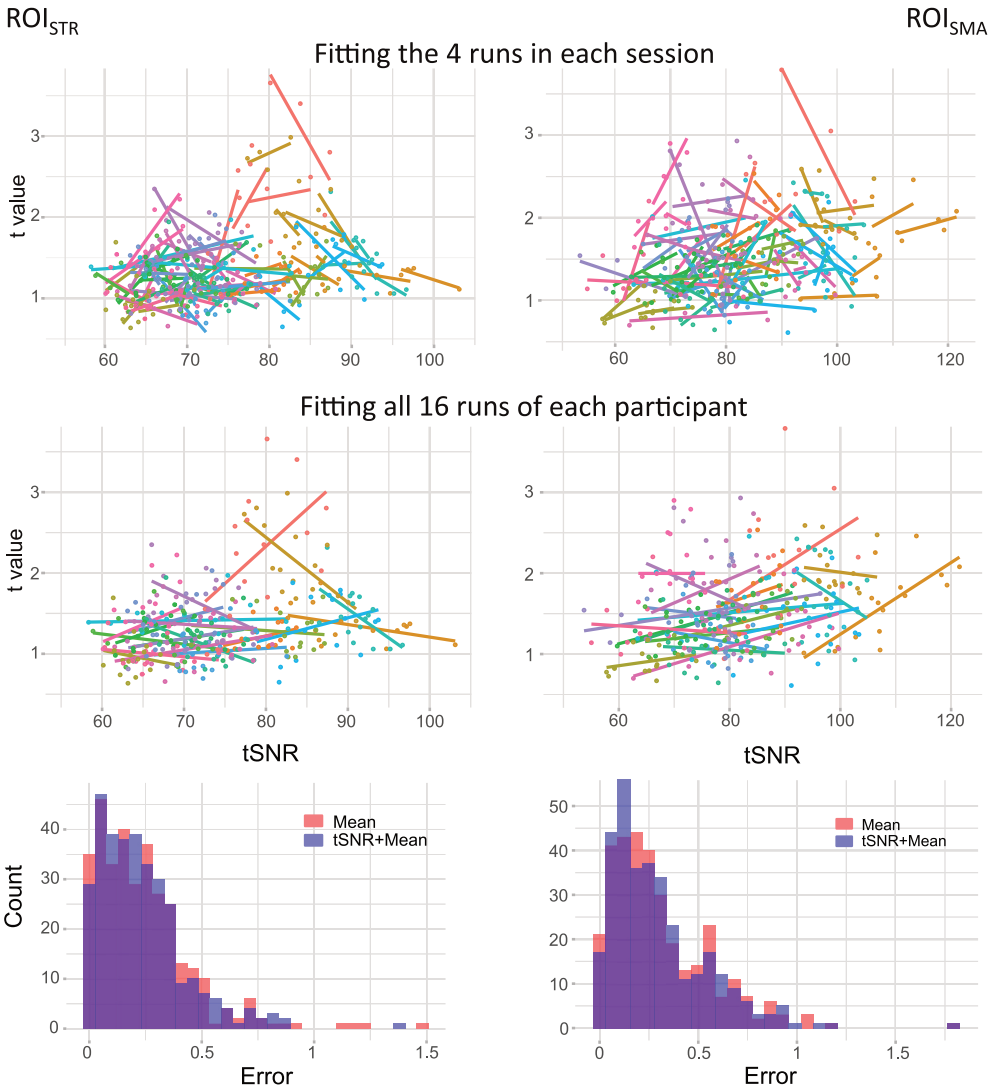


FIGURE 5 Correlations between mean temporal signal-to-noise ratio (tSNR) and mean t-value for each participant in two regions of interest (ROIs). Correlations are shown within both the striatum ROI (left) and the SMA ROI (right). Each point represents a functional magnetic resonance imaging (fMRI) run. The plots in rows 1 and 2 display linear fits (colored lines) across the 4 runs of each session for each participant, and all 16 runs for each participant, respectively. Each participant is shown in a different color. The histograms in the third row show the absolute error from two linear mixed effects models that seek to explain the variability in t-values as a function of tSNR plus participant-specific means (blue), or only the participant-specific means (red). The overlapping error distributions indicate that including tSNR does not significantly increase the model's ability to explain the variance in t-values across participants and task repetitions [Color figure can be viewed at wileyonlinelibrary.com]

tSNR was a good predictor of the resulting t-value on an individual level. Finally, we compared mixed-effects regression models that included tSNR as a predictor of t-values across all runs and participants to null models including only an intercept term with regard to the resulting absolute error between model-predicted and observed t-values (Figure 5, bottom row). Including tSNR as a predictor in the model did not significantly reduce the error in either the striatum or SMA ROIs ($t = 0.75$, $p = 0.45$; $t = 0.43$, $p = 0.67$, respectively).

3.5 | Other outcome measures

In both the SMA and the striatum ROIs, we found an unexpected, but consistent and statistically significant

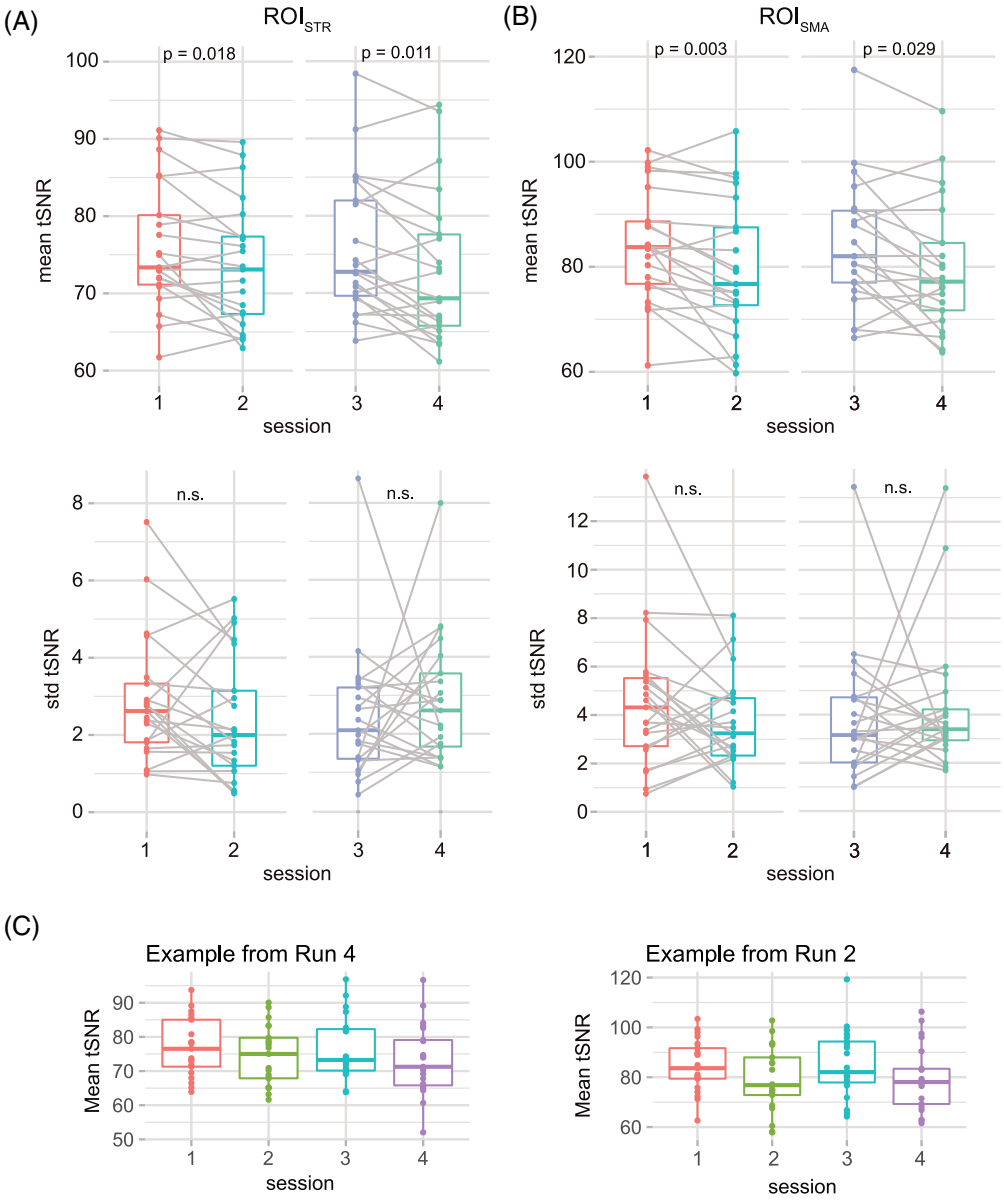


FIGURE 6 Differences across runs and sessions of the mean and standard deviation of temporal signal-to-noise ratio (tSNR). Each point represents either the mean (1st row) or the standard deviation (2nd row) across the four runs in each session of the region of interest (ROI)-averaged tSNR in the striatum (A) and the SMA (B) for a given participant. The four sessions are color coded only for visual guidance. The standard boxplots represent the median value as well as the 25th and 75th percentiles, while the whiskers include all the data that fall within the 95% confidence interval. The gray lines connect the data for each participant between the two sessions within each day (Sessions 1 and 2 on Day 1 and Sessions 3 and 4 on Day 2, respectively). The *p*-values are the result of two-tailed Wilcoxon tests and n.s. stands for nonsignificant. Note the drop of mean tSNR in Sessions 2 and 4 (i.e., the afternoon sessions). To ascertain robustness of this finding we investigated each run separately. Panel C provides Runs 2 and 4, where each point represents the mean of the ROI-averaged tSNR values across the 21 participants and the sessions are color coded only for visual guidance. The tSNR of the afternoon session was lower in every run (Runs 1 and 3 not shown) [Color figure can be viewed at wileyonlinelibrary.com]

decrease in the mean tSNR in the second session on each day (Figure 6(A,B)). This was true not only on average for the session but when each run within a given session was scrutinized separately (two of the four runs given in Figure 6(C)). This repeated pattern might be due to time of day effects[39] and may originate from the scanner (e.g., different temperature, the scanner warms-up by the second session compared to the first one) and/or the participant (e.g., physiological considerations or positioning). Specifically, in the DataSet_1_Group, the participants performed the task in two different satiety states, hungry (early morning), and fed (early afternoon).

Further, we found no significant association between tSNR and the possible confounds of age and BMI. Nor was there a significant difference in the variability of tSNR measures between the 2 days that would have indicated a longer-term change in daily scanner stability (Figure 7).

4 | DISCUSSION

We found that tSNR, which is widely used for evaluating the quality of data or efficacy of a given method, itself has poor reliability in vivo. More specifically, on repeated measures, the voxel-wise tSNR varied by up to 80%. With such high variability, in-vivo tSNR may not be a reliable outcome measure for testing incremental improvements in fMRI methods.

We alluded to a thought experiment in Section 1, in which tSNR was proposed as an outcome measure for optimizing the Ernst angle in an fMRI experiment with short TR. If the variability of tSNR on repeated scans can be up to 25% (see CoV in Figure 2), such an experiment would be challenging to carry out. It would require a group of 52 participants to identify even as large a difference as 10% in tSNR at a power of 0.80.

Applying such practical procedures, involving many other acquisition parameters and image processing variants,

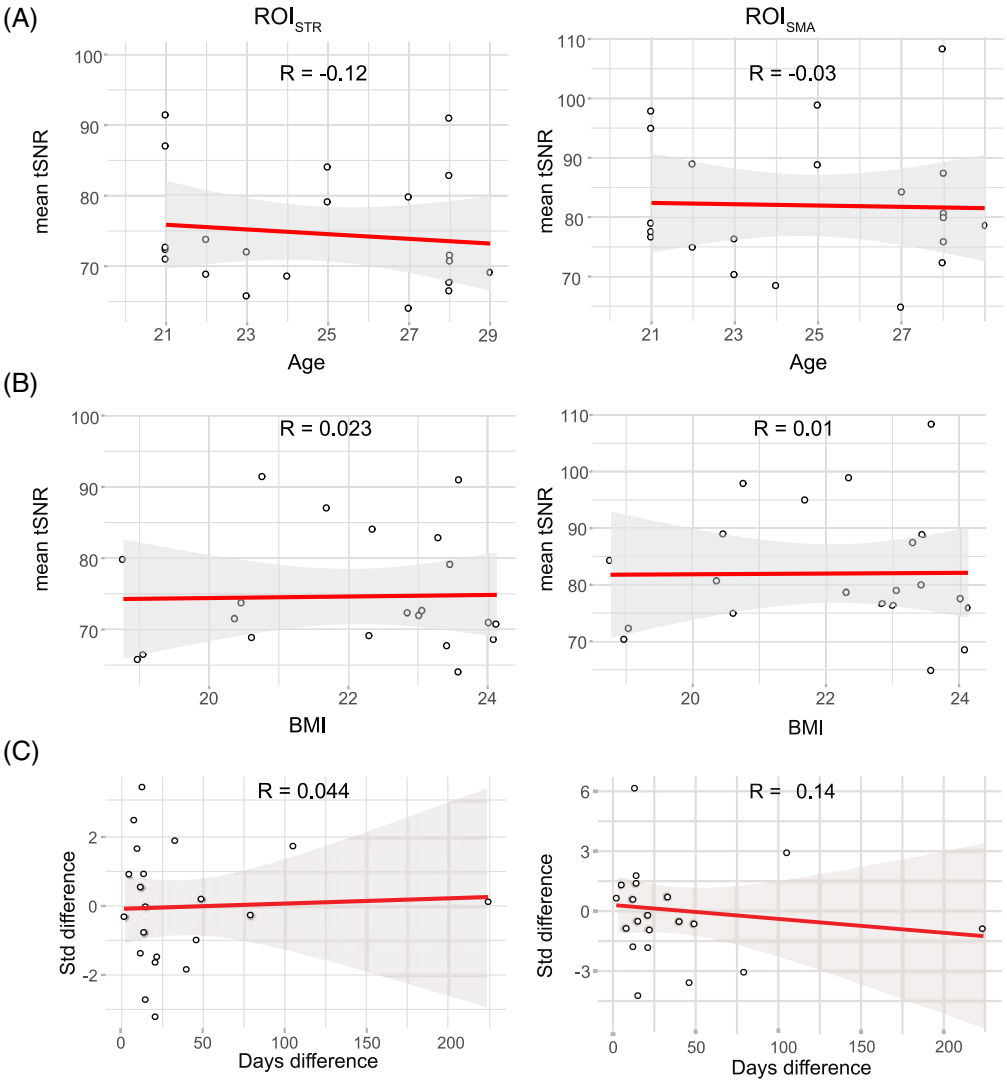


FIGURE 7 Correlations between the mean temporal signal-to-noise ratio (tSNR) and demographic or procedural variables. Pearson correlations between mean tSNR and age (A) as well as mean tSNR and body mass index (BMI) (B) for each participant averaging across sessions in both the striatum (left) and the SMA (right) regions of interest (ROIs). (C) Pearson correlations between the difference in days (day2–day1) and the difference in the standard deviation of tSNR values calculated for the eight runs on each day (std of tSNR day2–std of tSNR day1) [Color figure can be viewed at wileyonlinelibrary.com]

has been widespread in the effort to optimize fMRI experiments. Because these efforts often proceed without a handle on the underlying reproducibility of the outcome measure (e.g., tSNR), the reliability of these improvements is hard to infer. Even when a certain report involves a group of participants or repeated measures on a phantom, it is often the case that parameters are varied until an effect is found and that result is published with a recommendation for the final procedure used.[29] However, the reproducibility of these results in a separate group of participants or another set of measurements on phantoms is seldom investigated.

The findings in this report caution against such practices. Future efforts should factor the in vivo tSNR reproducibility level when designing studies or include empirical evidence that the planned experimental procedure (including scanner type, image processing pipeline, etc.) leads to lower variability in the voxel-wise in vivo tSNR.

Importantly, variability in results can originate from factors before data acquisition, including scanner choice,[40] interindividual variabilities,[41] for example, in head positioning in a multichannel coil with nonuniform SNR,[42] head motion,[43] and acquisition parameter settings.[44,45] Variation can arise even after the actual acquisition of data, with examples including choice of analysis software[46] and choices within the processing pipeline.[47] The present study encompasses many of these sources, but not all, and as such, it is likely that we report an underestimation of the total variability in tSNR.

This study does have some potential limitations that should be addressed in future work. First, DataSet_1_Group included only adult female participants, and thus our results are not necessarily representative of the general human population. The tSNR may be even less reliable in children, the elderly, or those suffering from neuropathologies. Second, it may be considered a limitation that tSNR maps were calculated before smoothing or corrections for autocorrelations as well as cardiac and respiratory signal variability, while the t-value maps were calculated after those processing steps. However, we deliberately proceeded in this way to capture signal variability from all sources. Smoothing would have artificially masked some of the variability in tSNR—a counter-productive exercise. Concerning physiological noise and autocorrelation, we must remember that we are not interested in tSNR per se. We are interested in its variability. However, even if our measure of tSNR is somewhat incorrect (i.e., physiological noise can decrease it, while autocorrelations can increase it somewhat) we would not expect an up to 80% difference on repeated measurements.[34] Furthermore, it is important to note that the TR of the time-series data acquisition was over 2 s; therefore, the impact of autocorrelations will be less severe.[48]

5 | CONCLUSIONS

We found that tSNR, extracted from repeated scans of the same individuals, is highly variable. Because of its wide use as a measure of data quality for fMRI studies and for evaluating incremental improvements in acquisition and processing methods, this variability in tSNR is disconcerting. With cognitive neuroscience focusing on small, even layer-specific, functional signals and medical approaches aiming to identify or treat disease in early stages based on subtle alterations, the voxel-wise reproducibility of tSNR is particularly relevant. We recommend incorporating the level of reproducibility of tSNR we report into the power calculations while planning future studies or to run similar prestudies to establish a more relevant measure for the local variability in tSNR at a given site.

ACKNOWLEDGMENTS

This work was supported by funding from the Zurich Center for Integrative Human Physiology, Philhuman Foundation, Foundation for Nutrition Research, Foundation for Scientific Research of the University of Zurich, Baugarten Foundation, and the Swiss NSF (grant Nos. PP00P1 150739 and 100014 165884 to P. N. T.). T. A. H. also acknowledges support from the European Union's Seventh Framework programme for research, technological development, and demonstration under grant agreement No. 607310 (Nudge-it) and Swiss NSF grant 32003B_166566. I. H. was supported by a grant from the Hungarian Brain Research Program (Nemzeti Agykutatási Program 2017-1.2.1-NKP-2017-00002). S. G. gratefully acknowledges the support of a Marlene Porsche Foundation scholarship for her PhD studies.

AUTHOR CONTRIBUTIONS

Susanna Gobbi, **Yoojin Lee**, and **Zoltan Nagy**: Conceived the study. **Susanna Gobbi**, **Zoltan Nagy**, and **István Homolya**: Collected the data. **Susanna Gobbi**, **Yoojin Lee**, and **Zoltan Nagy**: Analyzed the data. **Susanna Gobbi** and **Zoltan Nagy**: Wrote the paper. **Todd A. Hare** and **Philippe N. Tobler**: Provided guidance for the data analysis and contributed to the first draft of the paper.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available from the corresponding author upon reasonable request.

ORCID

Susanna Gobbi https://orcid.org/0000-0002-2115-3116

REFERENCES

1. Editorial. Unreliable research: trouble at the lab. *The Economist*. 2013;(19 Oct):1-10. http://www.economist.com/news/briefing/21588057-scientists-think-science-self-correcting-alarming-degree-it-not-trouble.
2. Munafò MR, Nosek BA, Bishop DVM, et al. A manifesto for reproducible science. *Nat Hum Behav*. 2017;1(1):1-9. https://doi.org/10.1038/s41562-016-0021
3. Baker M. Is there a reproducibility crisis? *Nature*. 2016;533:452-454.
4. Bennett CM, Miller MB. How reliable are the results from functional magnetic resonance imaging? *Ann N Y Acad Sci*. 2010;1191:133-155. https://doi.org/10.1111/j.1749-6632.2010.05446.x
5. Carp J. The secret lives of experiments: methods reporting in the fMRI literature. *Neuroimage*. 2012;63(1):289-300. https://doi.org/10.1016/j.neuroimage.2012.07.004
6. Editorial. Fostering reproducible fMRI research. *Nat Neurosci*. 2017;20(3):298-298. https://doi.org/10.1038/nn.4521
7. Elliott ML, Knodt AR, Ireland D, et al. What is the test-retest reliability of common task-functional MRI measures? New empirical evidence and a meta-analysis. *Psychol Sci*. 2020;31(7):792-806. https://doi.org/10.1177/0956797620916786
8. Nichols TE, Das S, Eickhoff SB, et al. Best practices in data analysis and sharing in neuroimaging using MRI. *Nat Neurosci*. 2017;20(3):299-303. https://doi.org/10.1038/nn.4500
9. Poldrack RA, Baker CI, Durnez J, et al. Scanning the horizon: towards transparent and reproducible neuroimaging research. *Nat Rev Neurosci*. 2017;18(2):115-126. https://doi.org/10.1038/nrn.2016.167
10. Casey BJ, Cohen JD, O'Craven K, et al. Reproducibility of fMRI results across four institutions using a spatial working memory task. *Neuroimage*. 1998;8(8):249-261.
11. Friedman L, Stern H, Brown GG, et al. Test-retest and between-site reliability in a multicenter fMRI study. *Hum Brain Mapp*. 2008;29(8):958-972. https://doi.org/10.1002/hbm.20440
12. Fröhner JH, Teckentrup V, Smolka MN, Kroemer NB. Addressing the reliability fallacy in fMRI: similar group effects may arise from unreliable individual effects. *Neuroimage*. 2019;195(March):174-189. https://doi.org/10.1016/j.neuroimage.2019.03.053
13. Maitra R, Roys SR, Gullapalli RP. Test-retest reliability estimation of functional MRI data. *Magn Reson Med*. 2002;48(1):62-70. https://doi.org/10.1002/mrm.10191
14. McGonigle DJ, Howseman AM, Athwal BS, Friston KJ, Frackowiak RSJ, Holmes AP. Variability in fMRI: an examination of intersession differences. *NeuroImage*. 2000;11(6 I):708-734. https://doi.org/10.1006/nimg.2000.0562
15. Plichta MM, Schwarz AJ, Grimm O, et al. Test-retest reliability of evoked BOLD signals from a cognitive-emotive fMRI test battery. *Neuroimage*. 2012;60(3):1746-1758. https://doi.org/10.1016/j.neuroimage.2012.01.129
16. Raemaekers M, du Plessis S, Ramsey NF, Weusten JMH, Vink M. Test-retest variability underlying fMRI measurements. *Neuroimage*. 2012;60(1):717-727. https://doi.org/10.1016/j.neuroimage.2011.11.061
17. Friston K. Ten ironic rules for non-statistical reviewers. *Neuroimage*. 2012;61(4):1300-1310. https://doi.org/10.1016/j.neuroimage.2012.04.018
18. Ingre M. Why small low-powered studies are worse than large high-powered studies and how to protect against "trivial" findings in research: comment on Friston (2012). *Neuroimage*. 2013;81:496-498. https://doi.org/10.1016/j.neuroimage.2013.03.030
19. Lindquist MA, Caffo B, Crainiceanu C. Ironing out the statistical wrinkles in "ten ironic rules". *Neuroimage*. 2013;81:499-502. https://doi.org/10.1016/j.neuroimage.2013.02.056
20. Poldrack RA, Fletcher PC, Henson RN, Worsley KJ, Brett M, Nichols TE. Guidelines for reporting an fMRI study. *Neuroimage*. 2008;40(2):409-414. https://doi.org/10.1016/j.neuroimage.2007.11.048
21. Ugurbil K, Toth L, Kim DS. How accurate is magnetic resonance imaging of brain function? *Trends Neurosci*. 2003;26(2):108-114. https://doi.org/10.1016/S0166-2236(02)00039-5
22. Friedman L, Glover GH. Report on a multicenter fMRI quality assurance protocol. *J Magn Reson Imaging*. 2006;23(6):827-839. https://doi.org/10.1002/jmri.20583
23. Weisskoff RM. Simple measurement of scanner stability for functional NMR imaging of activation in the brain. *Magn Reson Med*. 1996;36(4):643-645.
24. Triantafyllou C, Polimeni JR, Wald LL. Physiological noise and signal-to-noise ratio in fMRI with multi-channel array coils. *Neuroimage*. 2011;55(2):597-606. https://doi.org/10.1016/j.neuroimage.2010.11.084
25. Birn RM, Molloy EK, Patriat R, et al. The effect of scan length on the reliability of resting-state fMRI connectivity estimates. *Neuroimage*. 2013;83:550-558.
26. Lutti A, Thomas DL, Hutton C, Weiskopf N. High-resolution functional MRI at 3T: 3D/2D echo-planar imaging with optimized physiological noise correction. *Magn Reson Med*. 2013;69(6):1657-1664. https://doi.org/10.1002/mrm.24398
27. Poser BA, Versluis MJ, Hoogduin JM, Norris DG. BOLD contrast sensitivity enhancement and artifact reduction with multi-echo EPI: parallel-acquired inhomogeneity-desensitized fMRI. *Magn Reson Med*. 2006;55(6):1227-1235. https://doi.org/10.1002/mrm.20900
28. Feinberg DA, Setsompop K. Ultra-fast MRI of the human brain with simultaneous multi-slice imaging. *J Magn Reson*. 2013;229:90-100.
29. Button KS, Ioannidis JPA, Mokrysz C, et al. Power failure: why small sample size undermines the reliability of neuroscience. *Nat Rev Neurosci*. 2013;14(5):365-376. https://doi.org/10.1038/nrn3475
30. Chenevert TL, Malyarenko DI, Newitt D, et al. Errors in quantitative image analysis due to platform-dependent image scaling. *Transl Oncol*. 2014;7(1):65-71.
31. Gobbi S, Weber SC, Graf G, et al. Reduced neural satiety responses in women affected by obesity. *Neuroscience*. 2020;447:94-112. https://doi.org/10.1016/j.neuroscience.2020.07.022
32. Friston KJ, Ashburner JT, Kiebel SJ, Nichols TE, Penny WD. *Statistical Parametric Mapping: The Analysis of Functional Brain Images*. London: Elsevier; 2007.
33. Ashburner J, Friston KJ. Unified segmentation. *Neuroimage*. 2005;26(3):839-851.
34. Kasper L, Bollmann S, Diaconescu AO, et al. The PhysIO toolbox for modeling physiological noise in fMRI data. *J Neurosci Methods*. 2017;276:56-72. https://doi.org/10.1016/j.jneumeth.2016.10.019

35. Rolls ET, Huang CC, Lin CP, Feng J, Joliot M. Automated anatomical labelling atlas 3. *Neuroimage*. 2020;206(May 2019):116189. https://doi.org/10.1016/j.neuroimage.2019.116189
36. Carnell S, Gibson C, Benson L, Ochner CN, Geliebter A. Neuroimaging and obesity: current knowledge and future directions. *Obes Rev*. 2012;13(1):43-56. https://doi.org/10.1111/j.1467-789X.2011.00927.x
37. Kringelbach ML, Stein A. Cortical mechanisms of human eating. *Frontiers in Eating and Weight Regulation*. Basel: Karger; 2009:164-175. https://doi.org/10.1159/000264404
38. Bates D, Mächler M, Bolker B, Walker S. Fitting linear mixed-effects models using lme4. *J Stat Softw*. 2015;67(1). https://doi.org/10.18637/jss.v067.i01
39. Orban C, Kong R, Li J, Chee MWL, Yeo BTT. Time of day is associated with paradoxical reductions in global signal fluctuation and functional connectivity. *PLoS Biol*. 2020;18(2):e3000602. https://doi.org/10.1371/journal.pbio.3000602
40. Lee Y, Callaghan MF, Acosta-Cabronero J, Lutti A, Nagy Z. Establishing intra- and inter-vendor reproducibility of T1 relaxation time measurements with 3T MRI. *Magn Reson Med*. 2019;81(1):454-465. https://doi.org/10.1002/mrm.27421
41. Kirilina E, Lutti A, Poser BA, Blankenburg F, Weiskopf N. The quest for the best: the impact of different EPI sequences on the sensitivity of random effect fMRI group analyses. *Neuroimage*. 2016;126:49-59. https://doi.org/10.1016/j.neuroimage.2015.10.071
42. Howseman AM, McGonigle DJ, Grootoonk S, et al. Assessment of the Variability in fMRI Data Sets Due to Subject Positioning and Calibration of the MRI Scanner. In *Proceedings of the 4th Annual Meeting of OHBM*, *Montreal*, *Canada* (p. Neuroimage 7, S599). Montreal, Canada: NeuroImage. 1998.
43. Liu TT. Noise contributions to the fMRI signal: an overview. *Neuroimage*. 2016;143:141-151. https://doi.org/10.1016/j.neuroimage.2016.09.008
44. Deichmann R, Gottfried JA, Hutton C, Turner R. Optimized EPI for fMRI studies of the orbitofrontal cortex. *Neuroimage*. 2003;19:430-441.
45. Wald LL, Polimeni JR. Impacting the effect of fMRI noise through hardware and acquisition choices—implications for controlling false positive rates. *Neuroimage*. 2017;154(December 2016):15-22. https://doi.org/10.1016/j.neuroimage.2016.12.057
46. Soares JM, Magalhães R, Moreira PS, et al. A Hitchhiker's guide to functional magnetic resonance imaging. *Front Neurosci*. 2016;10:515. https://doi.org/10.3389/fnins.2016.00515
47. Cusack R, Brett M, Osswald K. An evaluation of the use of magnetic field maps to undistort echo-planar images. *Neuroimage*. 2003;18(1):127-142.
48. Corbin N, Todd N, Friston KJ, Callaghan MF. Accurate modeling of temporal correlations in rapidly sampled fMRI time series. *Hum Brain Mapp*. 2018;39(10):3884-3897. https://doi.org/10.1002/hbm.24218

How to cite this article: Gobbi S, Lee Y, Homolya I, Tobler PN, Hare TA, Nagy Z. On the reproducibility of in vivo temporal signal-to-noise ratio and its utility as a predictor of subject-level t-values in a functional magnetic resonance imaging study. *Int J Imaging Syst Technol*. 2021;1-12. https://doi.org/10.1002/ima.22617